كلية الحاسبات والذكاء الإصطناعي

BS111
Probability and Statistics

Lecture 12

Dr. Ahmed Hagag

Faculty of Computers and Artificial Intelligence
Benha University

2021

Ch 4: Descriptive Statistics

Data Description (1/2)

- Population and Sample in a Statistical Study.
- Types of Data.
- Representation of Data.
- Numerical Summaries of Quantitative Data.
  - Measures of Centrality (Central Tendency/location)
  - Measures of Variability (Dispersion/Spread)
  - Measures of Shape. (Skewness, Kurtosis)
  - Measures of Relative Position. (Percentiles, Quartiles, IQR, Coefficient of Variation).

Data Description (2/2)

- Stem-and-Leaf Diagrams.
- Box-Whisker Plot (Box Plot).
- Time Sequence Plots.
- Scatter Diagrams.

Population and Sample (1/3)

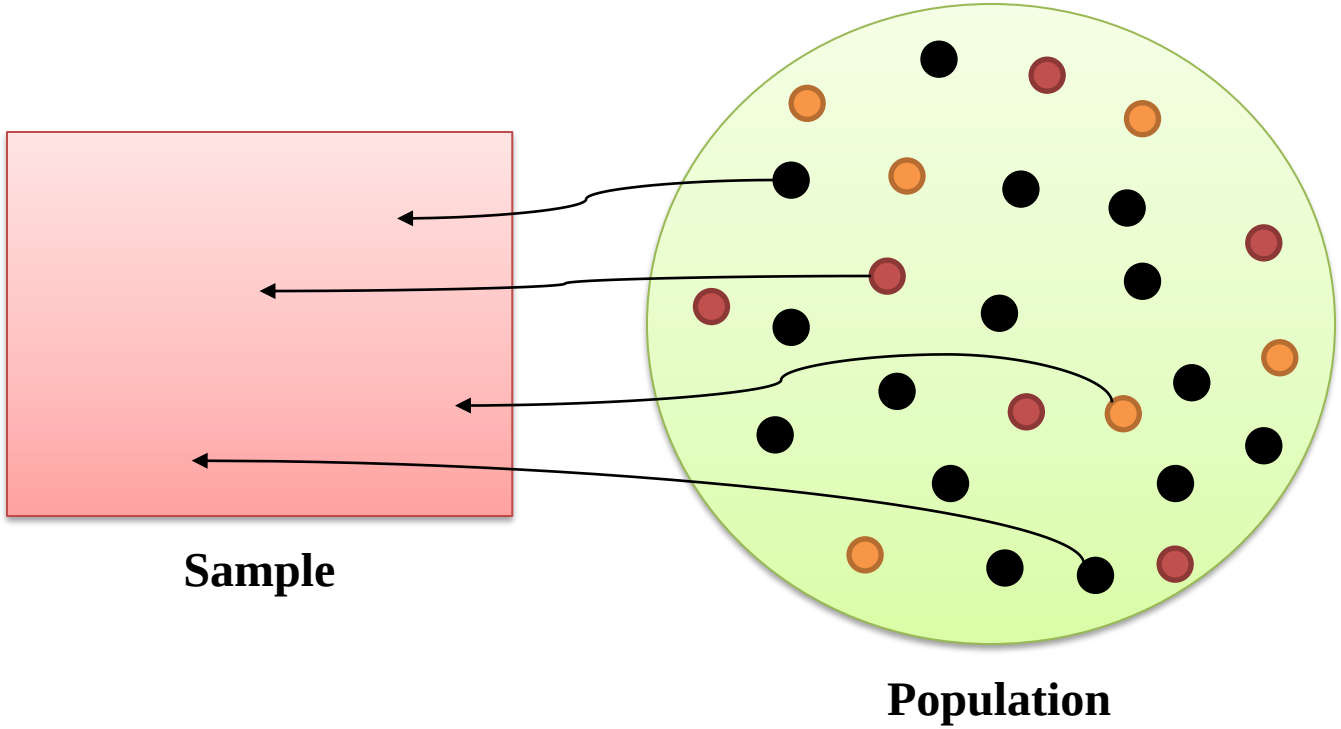
Definitions (1/9):

- A **population** is a collection of **all** elements that possess a characteristic of interest.
- Populations can be *finite* or *infinite*. A population where all the elements are easily countable may be considered as finite, and a population where all the elements are not easily countable as infinite.

Population and Sample (1/3)

Definitions (2/9):

- A **portion** of a population selected for study is called a **sample**.



Population and Sample (1/3)

Definitions (3/9):

First type is called **biased samples**. When one or more parts of the population are favored over others.

- Convenience Sample.
- Voluntary Responses Sample.

Population and Sample (1/3)

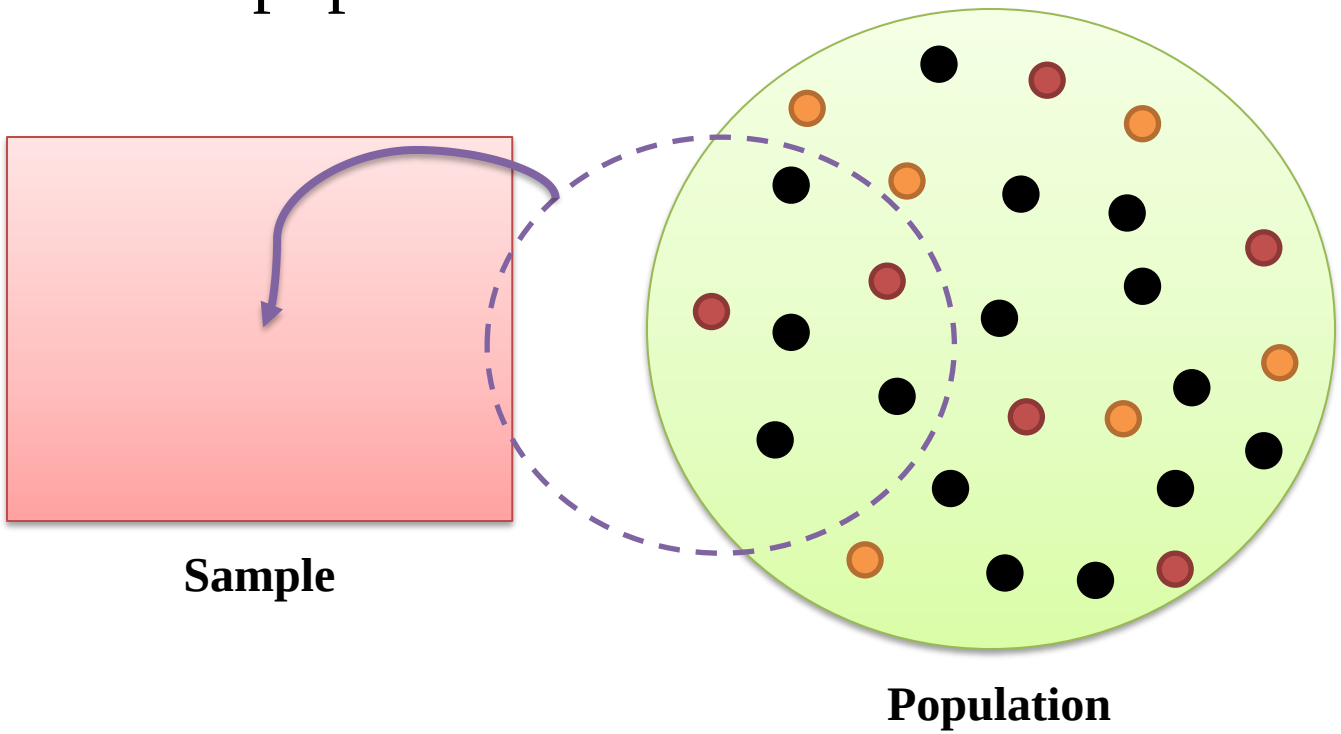
Definitions (4/9):

***Convenience Sample*:** only include elements that are easy to reach from the population.

Population and Sample (1/3)

Definitions (4/9):

***Convenience Sample*:** only include elements that are easy to reach from the population.



Population and Sample (1/3)

Definitions (5/9):

***Voluntary Responses Sample*:** consist of people that have chosen to include themselves (e.g., survey: people with a strong interest for the survey topic are the ones who are most likely to respond).

Population and Sample (1/3)

Definitions (6/9):

In order to avoid having bias, we want our sample to be random, called **unbiased samples**.

- Simple Random Sample (SRS).
- Stratified Random Sample.
- Clustering (Multistage) Random Sample.

Population and Sample (1/3)

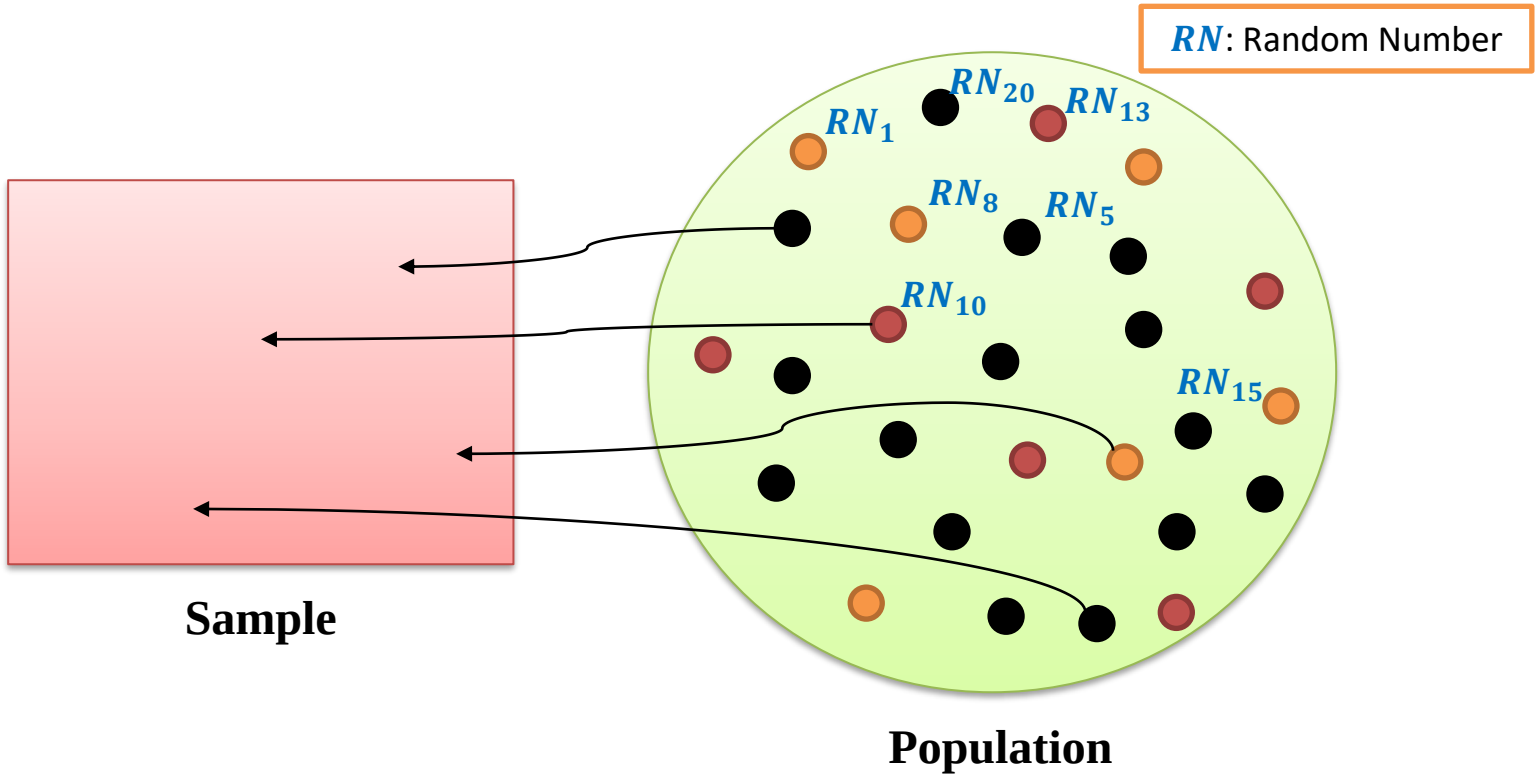
Definitions (7/9):

***Simple Random Sample (SRS)*:** if each element of the population has the same chance of being included in the sample.

Population and Sample (1/3)

Definitions (7/9):

***Simple Random Sample (SRS)*:** if each element of the population has the same chance of being included in the sample.



Population and Sample (1/3)

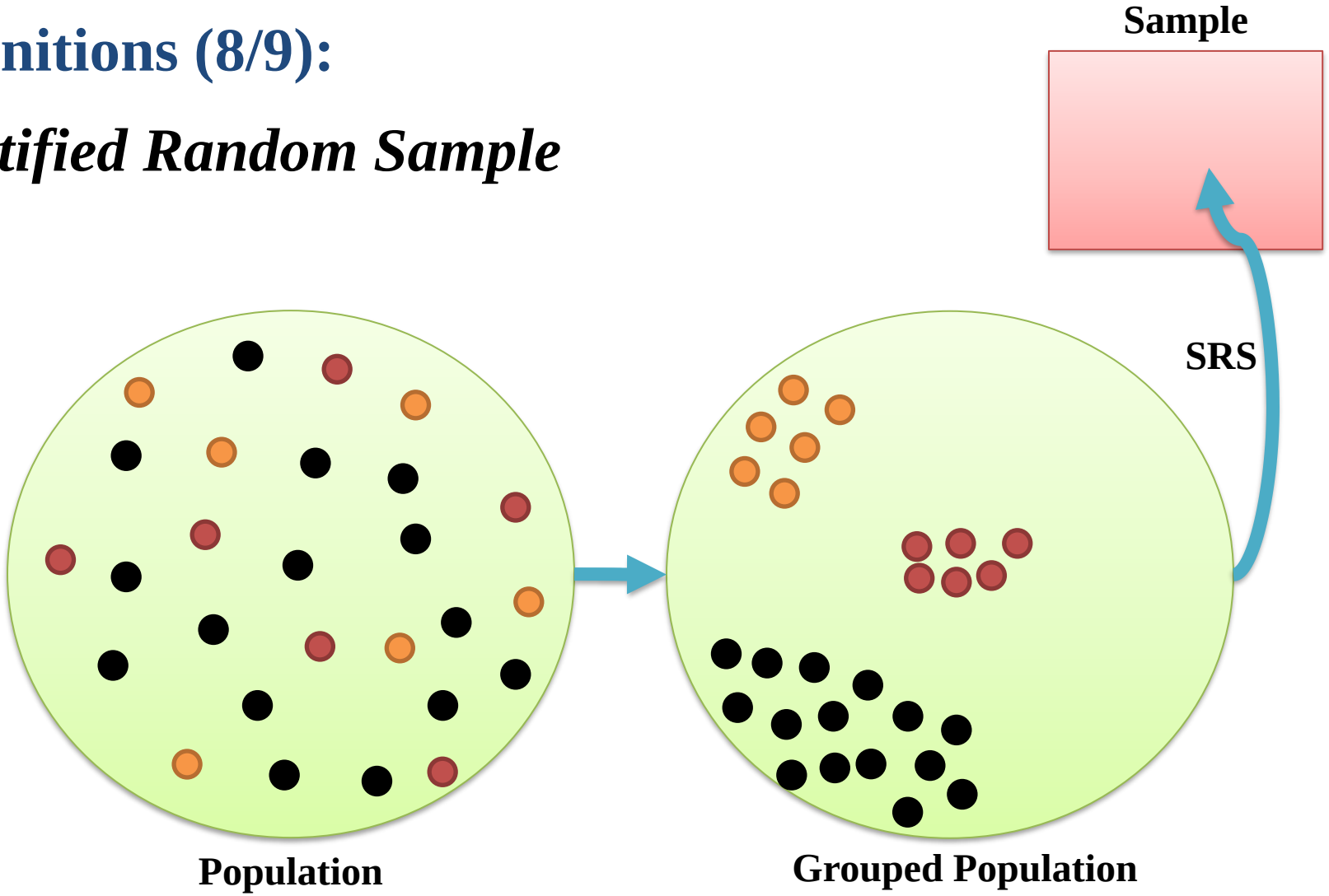
Definitions (8/9):

***Stratified Random Sample*:** elements of the population are divided into groups represent the similar types (called Stratum). Then, The SRS is taken within each stratum to select some items and combine the results to get the sample.

Population and Sample (1/3)

Definitions (8/9):

Stratified Random Sample



Population and Sample (1/3)

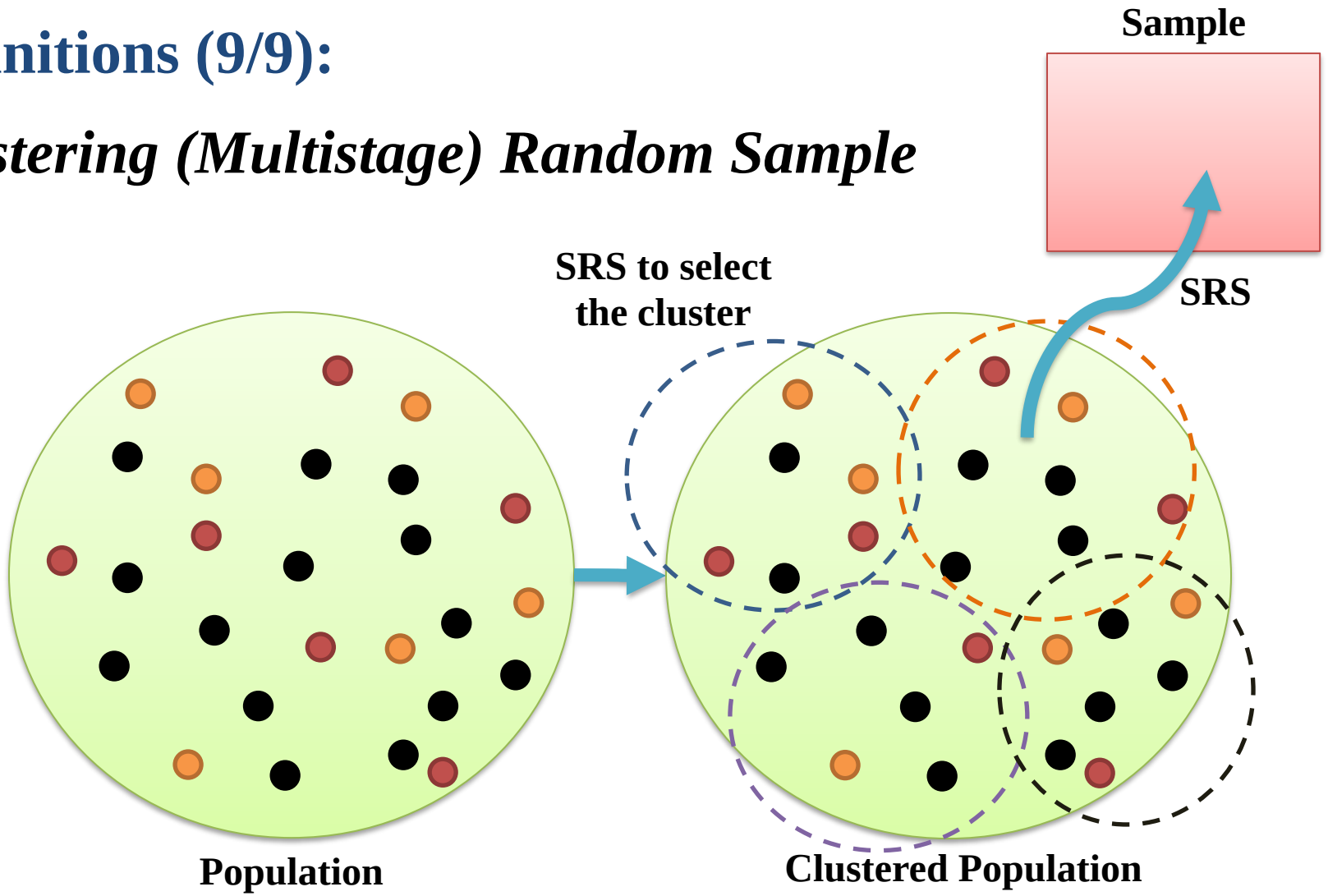
Definitions (9/9):

***Clustering (Multistage) Random Sample*:** elements of the population are divided into different clusters. Then, the SRS is taken to select one cluster. After that, the SRS is taken again within the selected cluster to select some items.

Population and Sample (1/3)

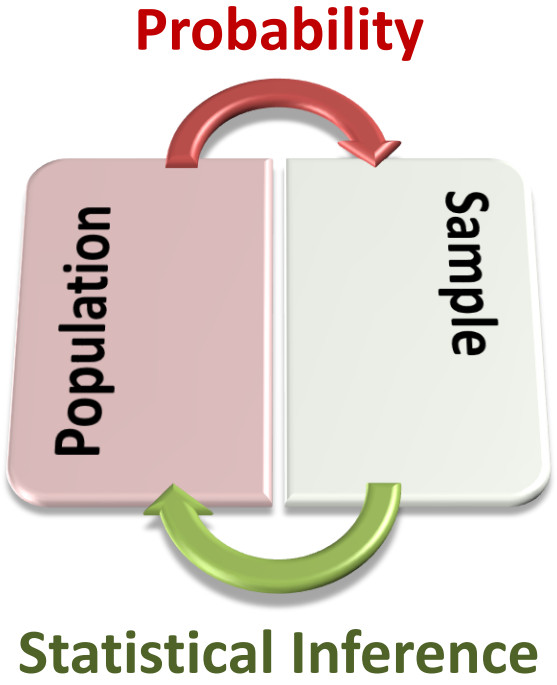
Definitions (9/9):

Clustering (Multistage) Random Sample



Population and Sample (2/3)

Probability And Inferential Statistics:

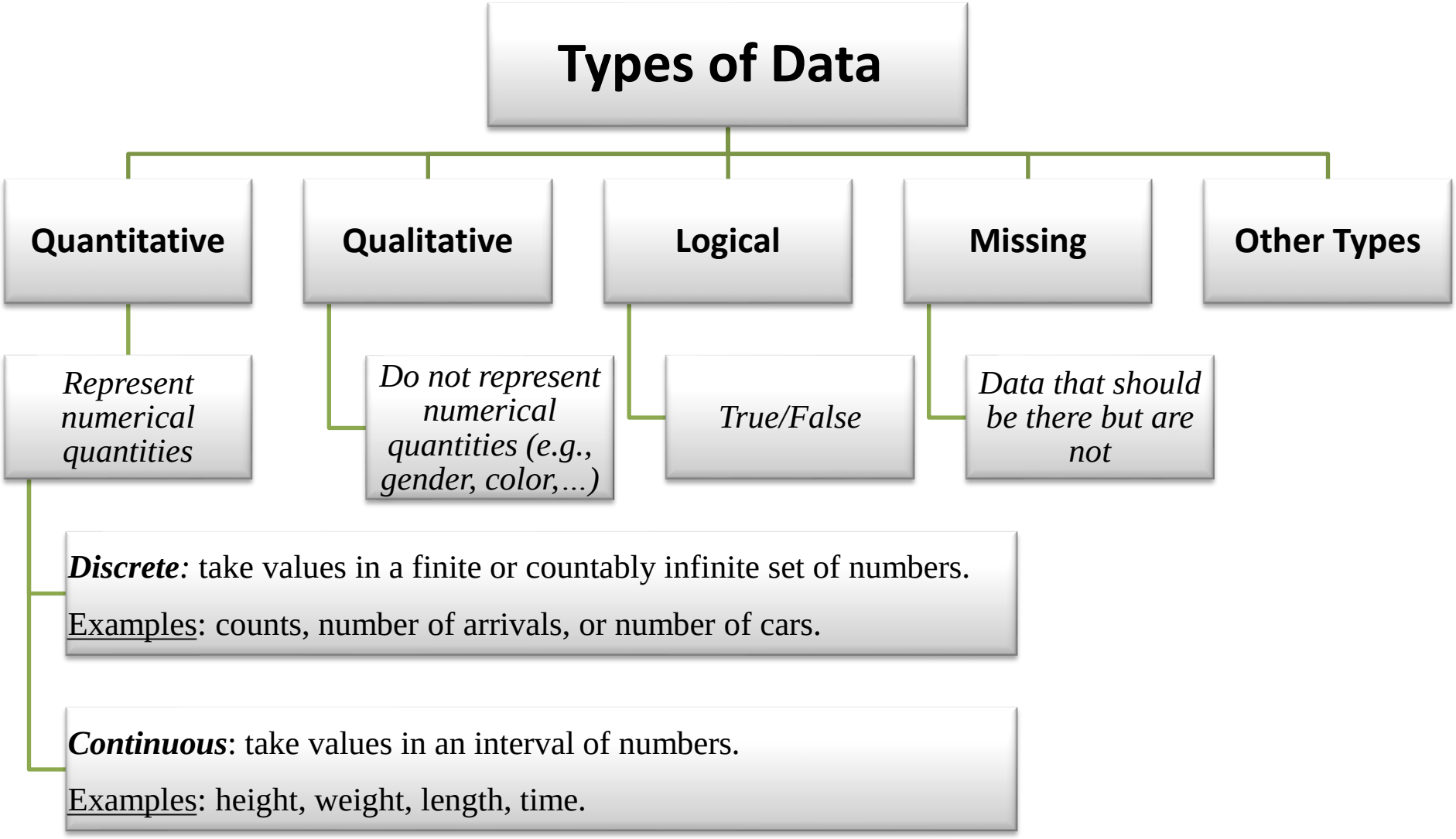


Population and Sample (3/3)

Statistics:

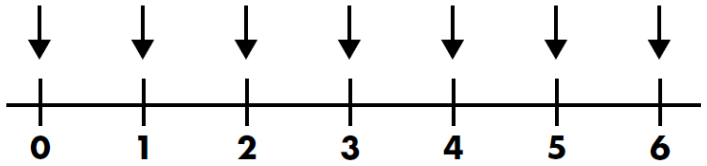
- Statistics concerns data; their collection, analysis, and interpretation.
- Descriptive statistics concerns the summarization of data. We have a dataset and we would like to describe the data set in multiple ways.
- Descriptive statistics consists of the collection, organization, summarization, and presentation of data.

Types of Data (1/6)


Types of Data
Quantitative
Qualitative
Logical
Missing
Other Types
Represent numerical quantities
Do not represent numerical quantities (e.g., gender, color,…)
True/False
Data that should be there but are not
Discrete: take values in a finite or countably infinite set of numbers.
Examples: counts, number of arrivals, or number of cars.
Continuous: take values in an interval of numbers.
Examples: height, weight, length, time.

Types of Data (2/6)

Discrete Quantitative Data:



Example:

12 people were asked about their own cars and the results were recorded as follows:

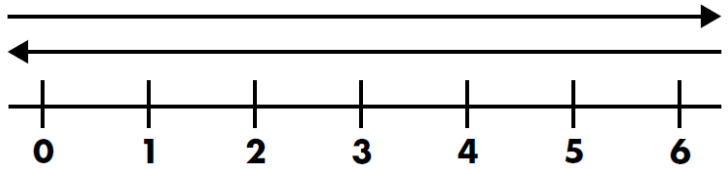
2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Car(s)

The measuring: *numerical*.

Types of Data (3/6)

Continuous Quantitative Data:



Example:

10 people were asked about their height and the results were recorded as follows:

$$175, 184, 160, 170, 173, 170, 175, 165, 171, 176$$

cm

The measuring: *numerical*.

Types of Data (4/6)

Qualitative Data:

Example:

11 people were asked about their gender and the results were recorded as follows:

$$M, F, M, M, M, F, F, M, M, M, M$$

The measuring: *nominal*.

Types of Data (5/6)

Logical Data:

The value is either TRUE or FALSE (note that equivalently you can use 1 = TRUE, 0 = FALSE).

Example:

Using R

```
> x <- 5:9
> y <- (x < 7.3)
> y
[1]  TRUE  TRUE  TRUE  FALSE  FALSE
```

Types of Data (6/6)

Missing Data:

Data that should be there but are not. **R** reserves the special symbol `NA` to representing missing data.

Example:

Using R

```
> x <- c(3, 7, NA, 4, 7)
> y <- c(5, NA, 1, 2, 2)
> x + y
[1]  8  NA  NA  6  9
> sum(x, na.rm = TRUE)
[1]  21
```

Revision Questions

Determine whether each statement is true or false:

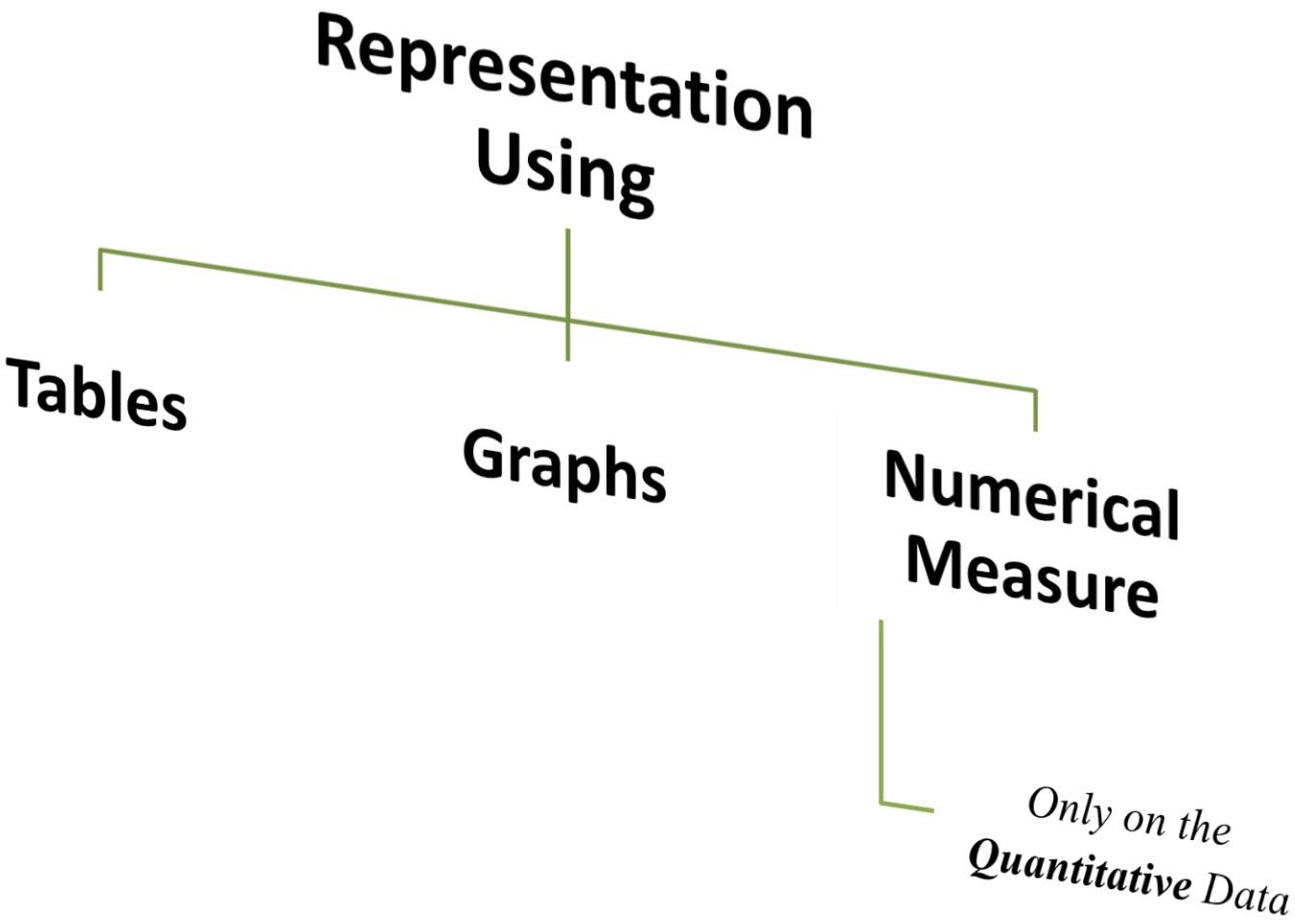
1. A sample consists of all subjects that are being studies. ( )
2. Descriptive statistics consists of the collection, organization, summarization, and presentation of data. ( )
3. The number of absences per year that a worker has is an example a continuous variable. ( )
4. The variable age is an example of discrete variable. ( )
5. Data that can be classified according to color are measured as nominal. ( )

Revision Questions

Determine whether each statement is true or false:

1. A sample consists of all subjects that are being studies. (F) (population)
2. Descriptive statistics consists of the collection, organization, summarization, and presentation of data. (T)
3. The number of absences per year that a worker has is an example a continuous variable. (F) (discrete)
4. The variable age is an example of discrete variable. (F) (continuous)
5. Data that can be classified according to color are measured as nominal. (T)

Representation of Data (1/8)



Representation of Data (2/8)

Example1: (1/6)

11 people were asked about their gender and the results were recorded as follows:

$$M, F, M, M, M, F, F, M, M, M, M$$

The type of this data raw is: ***Qualitative Data***

Representation of Data (2/8)

Example1: (2/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
M			
F			
Sum			

Representation of Data (2/8)

Example1: (2/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
M	𝍸 ///	8	8/11
F	///	3	3/11
Sum	𝍸 𝍸 /	**11**	**1**

Representation of Data (2/8)

Example1: (2/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
M	𝍸 ///	8	8/11
F	///	3	3/11
Sum	𝍸 𝍸 /	**11**	**1**

sample size

Representation of Data (2/8)

Example1: (3/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
M	𝍸 ///	8	8/11	
F	///	3	3/11	
Sum	𝍸 𝍸 /	**11**	**1**	

Representation of Data (2/8)

Example1: (3/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
M	𝍸 ///	8	8/11	72.73%
F	///	3	3/11	27.27%
Sum	𝍸 𝍸 /	**11**	**1**	**100%**

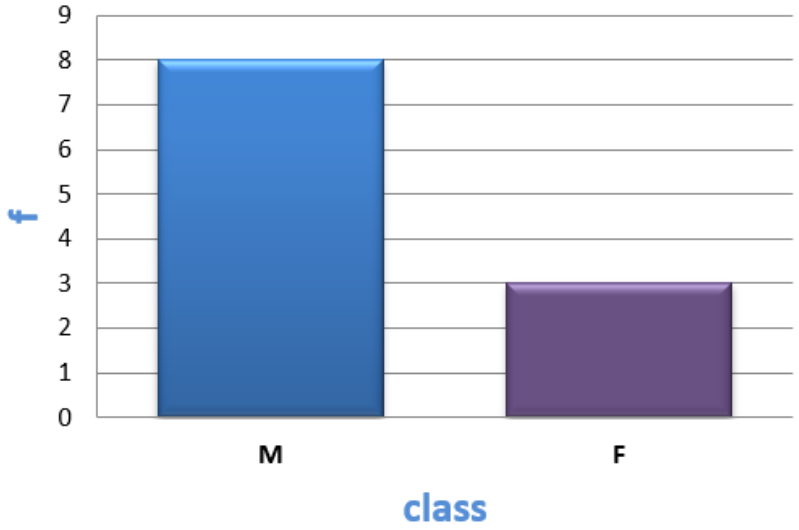
Representation of Data (2/8)

Example1: (4/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Bar Graphs

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
M	𝍸 ///	8	8/11	72.73%
F	///	3	3/11	27.27%
Sum	𝍸 𝍸 /	11	1	100%



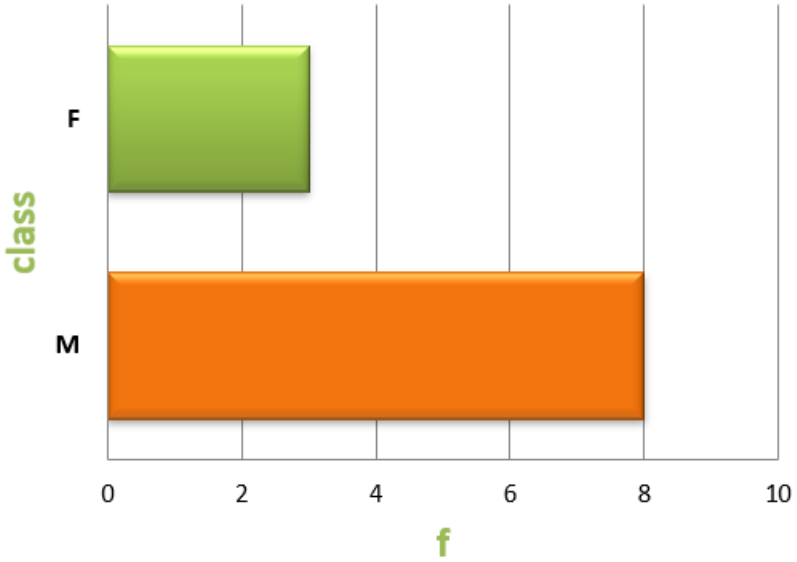
Representation of Data (2/8)

Example1: (5/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Bar Graphs

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
M	𝍸 ///	8	8/11	72.73%
F	///	3	3/11	27.27%
Sum	𝍸 𝍸 /	11	1	100%



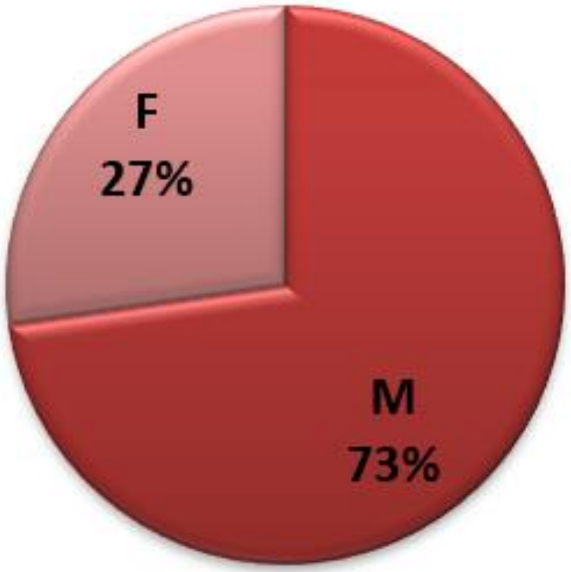
Representation of Data (2/8)

Example1: (6/6)

$$M, F, M, M, M, F, F, M, M, M, M$$

Using Pie Charts

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
M	𝍸 ///	8	8/11	72.73%
F	///	3	3/11	27.27%
Sum	𝍸 𝍸 /	11	1	100%



Angle of a slice = (Relative frequency of the given class) × 360

Ex. Angle of a slice for class M = 8/11 * 360 = 261.81°

Representation of Data (3/8)

Example2: (1/4)

A study of 552 first-year college students asked about their preferences for online resources. One question asked them to pick their favorite. Here are the results

Resource	Count (n)
Google or Google Scholar	406
Library database or website	75
Wikipedia or online encyclopedia	52
Other	19
Total	552

The type of this data raw is: ***Qualitative Data***

Representation of Data (3/8)

Example2: (2/4)

Resource	Count (n)
Google or Google Scholar	406
Library database or website	75
Wikipedia or online encyclopedia	52
Other	19
Total	552

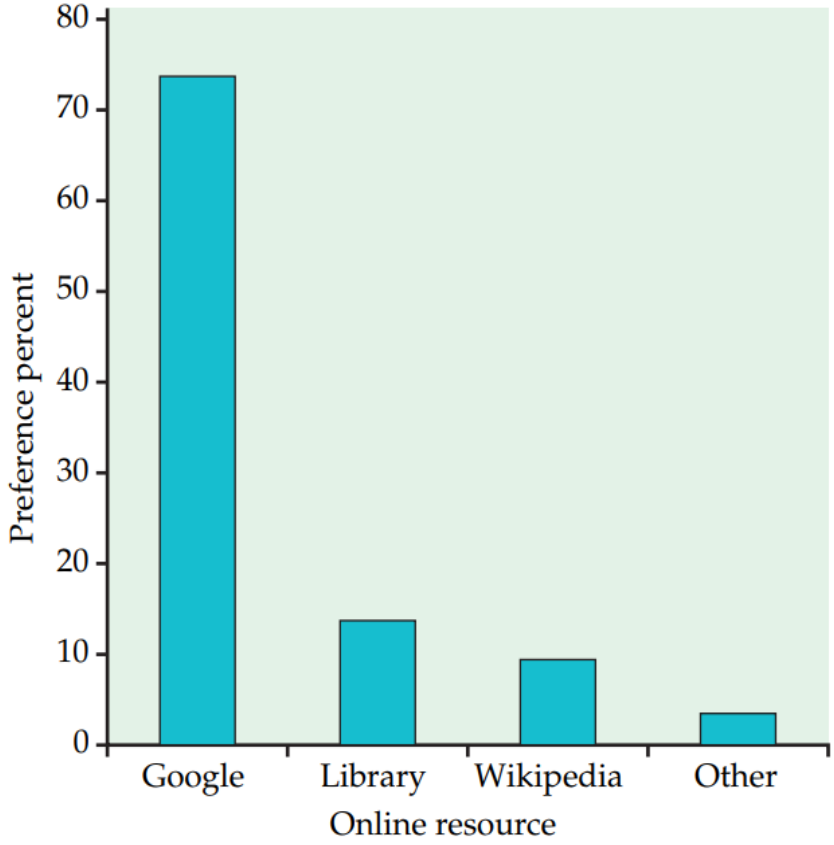
Resource	Percent (%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0

Representation of Data (3/8)

Example2: (3/4)

Using Bar Graphs

Resource	Percent (%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0

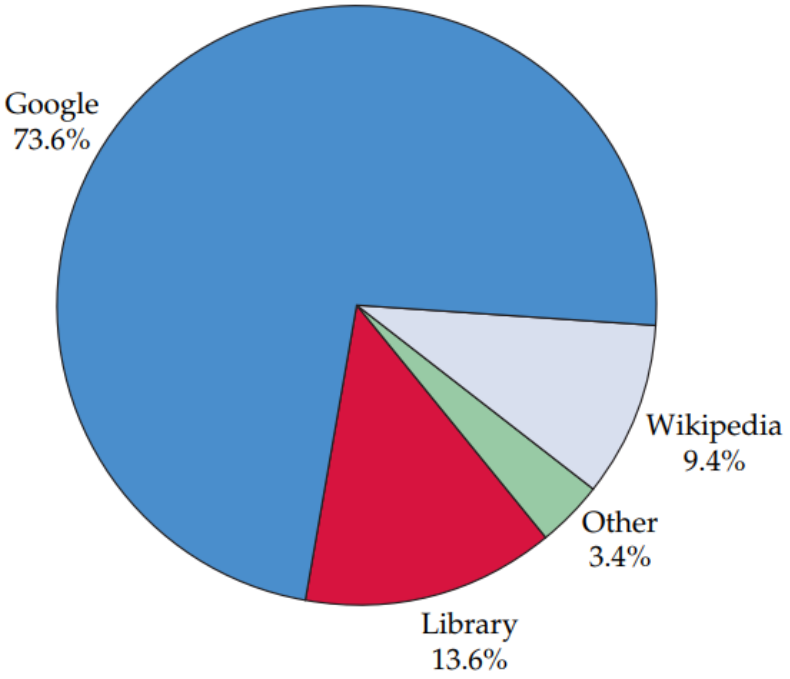


Representation of Data (3/8)

Example2: (4/4)

Using Pie Charts

Resource	Percent (%)
Google or Google Scholar	73.6
Library database or website	13.6
Wikipedia or online encyclopedia	9.4
Other	3.4
Total	100.0



Angle of a slice = (Percent%) × 3.6

Ex. Angle of a slice for class Wikipedia = 9.4* 3.6 = 33.84°

Representation of Data (4/8)

Example3: (1/7)

12 people were asked about their own cars and the results were recorded as follows:

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Car(s)

The type of this data raw is: ***Discrete Quantitative Data***

Representation of Data (4/8)

Example3: (2/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
0			
2			
3			
4			
Sum			

Representation of Data (4/8)

Example3: (2/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
0	/	1	1/12
2	𝍸///	8	8/12
3	/	1	1/12
4	//	2	2/12
Sum	𝍸 𝍸	12	1

sample size

Representation of Data (4/8)

Example3: (3/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Frequency Table (Grouped Data)

Called: *Frequency distribution for the data*.

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
0	/	1	1/12	08.33 %
2	𝍸///	8	8/12	66.67 %
3	/	1	1/12	08.33 %
4	//	2	2/12	16.67 %
Sum	𝍸 𝍸	12	1	100 %

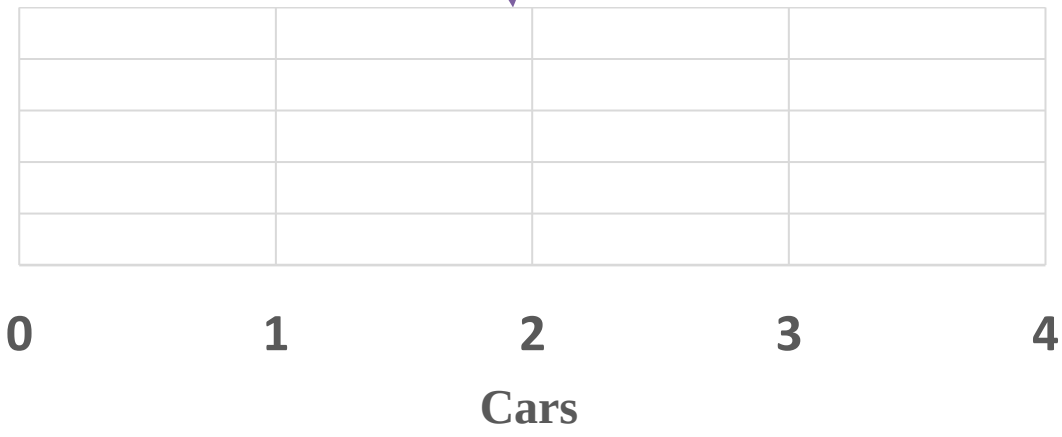
Representation of Data (4/8)

Example3: (4/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Dot Plot

A dot plot is one of the simplest graphs.



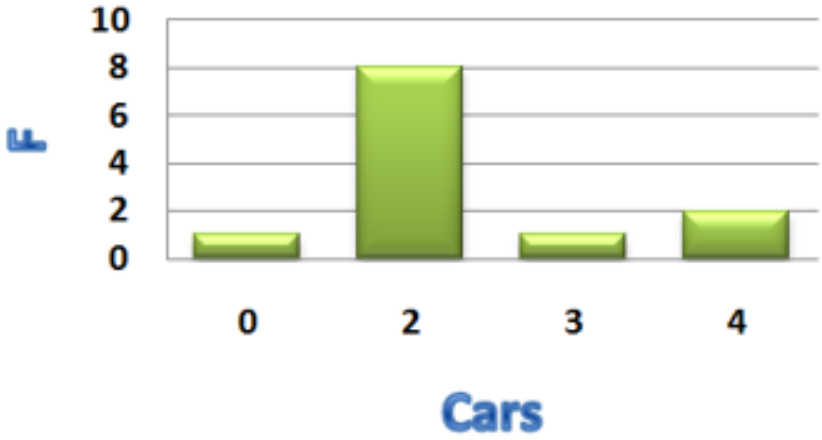
Representation of Data (4/8)

Example3: (5/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Bar Graphs

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
0	/	1	1/12	08.33 %
2	𝍸///	8	8/12	66.67 %
3	/	1	1/12	08.33 %
4	//	2	2/12	16.67 %
Sum	𝍸𝍸	12	1	100 %



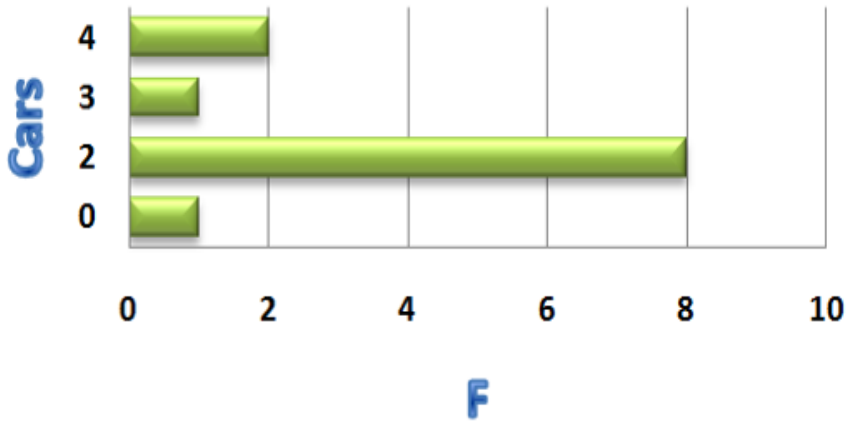
Representation of Data (4/8)

Example3: (6/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Bar Graphs

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
0	/	1	1/12	08.33 %
2	𝍸///	8	8/12	66.67 %
3	/	1	1/12	08.33 %
4	//	2	2/12	16.67 %
Sum	𝍸𝍸	12	1	100 %



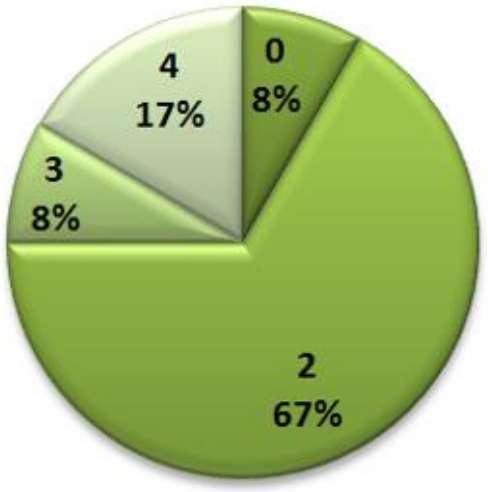
Representation of Data (4/8)

Example3: (7/7)

2, 0, 4, 2, 2, 3, 2, 2, 4, 2, 2, 2

Using Pie Charts

Class	Tally	Frequency (F)	Relative Frequency (R.F.)	Percentage
0	/	1	1/12	08.33 %
2	𝍸///	8	8/12	66.67 %
3	/	1	1/12	08.33 %
4	//	2	2/12	16.67 %
Sum	𝍸 𝍸	12	1	100 %



Angle of a slice = (Relative frequency of the given class) × 360

Ex. Angle of a slice for class 0 = 1/12 * 360 = 30°

Representation of Data (5/8)

Example4: (1/2)

The following data give the number of defective motors received in 20 different shipments:

8	12	10	16	10	25	21	15	17	5
26	21	29	8	6	21	10	17	15	13

Construct a dot plot for these data.

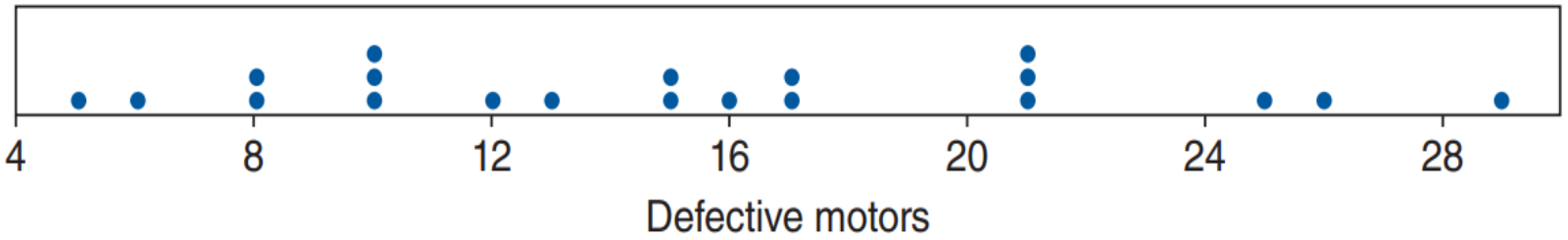
The type of this data raw is: ***Discrete Quantitative Data***

Representation of Data (5/8)

Example4: (2/2)

8	12	10	16	10	25	21	15	17	5
26	21	29	8	6	21	10	17	15	13

Dot Plot



Representation of Data (6/8)

Example5: (1/9)

10 people were asked about their height and the results were recorded as follows:

$$175, 184, 160, 170, 173, 170, 175, 165, 171, 176$$

cm

The type of this data raw is: ***Continuous Quantitative Data***

Representation of Data (6/8)

Example5: (2/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Because the type of this data is continuous quantitative, we follow the following steps:

1. Find the range R of the data that is defined as:

$$\textbf{Range} = R = \text{largest data point} - \text{smallest data point}$$

Representation of Data (6/8)

Example5: (3/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

2. Divide the data set into an appropriate number of ***classes***. The classes are also sometimes called *intervals*, *categories*, *cells*, or *bins*.

The number of classes is k. There is no "best" number of classes. However, Sturges's formula is often used, given by:

$$\text{Number of classes} = k = \lfloor 1 + 3.322 \log n \rfloor$$

Or we can use a simple formula as follows:

$$k = \sqrt{n}$$

where n is the total number of data points in a given data set.

Representation of Data (6/8)

Example5: (4/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

3. Determine the width of classes as follows:

Class width = Range / Number of Classes = R/k

4. Finally, preparing the frequency distribution table is achieved by assigning each data point to an appropriate class.

Representation of Data (6/8)

Example5: (5/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

1. Range $= R = 184 - 160 = 24$

2. The number of classes $k = [1 + 3.322 \log 10\ ] = 4$

3. Class width $= R/k = 24/4 = 6$

4. The four classes used to prepare the frequency distribution table are as follows:

$$[160 - 166), [166 - 172), [172 - 178), [178 - 184]$$

Representation of Data (6/8)

Example5: (6/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Using Frequency Table (Grouped Data)

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
[160 – 166)			
[166 – 172)			
[172 – 178)			
[178 – 184]			
Sum			

Representation of Data (6/8)

Example5: (6/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Using Frequency Table (Grouped Data)

Class	Tally	Frequency (F)	Relative Frequency (R.F.)
[160 – 166)	//	2	2/10
[166 – 172)	///	3	3/10
[172 – 178)	////	4	4/10
[178 – 184]	/	1	1/10
Sum	////-////	**10**	**1**

Representation of Data (6/8)

Example5: (7/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Histograms

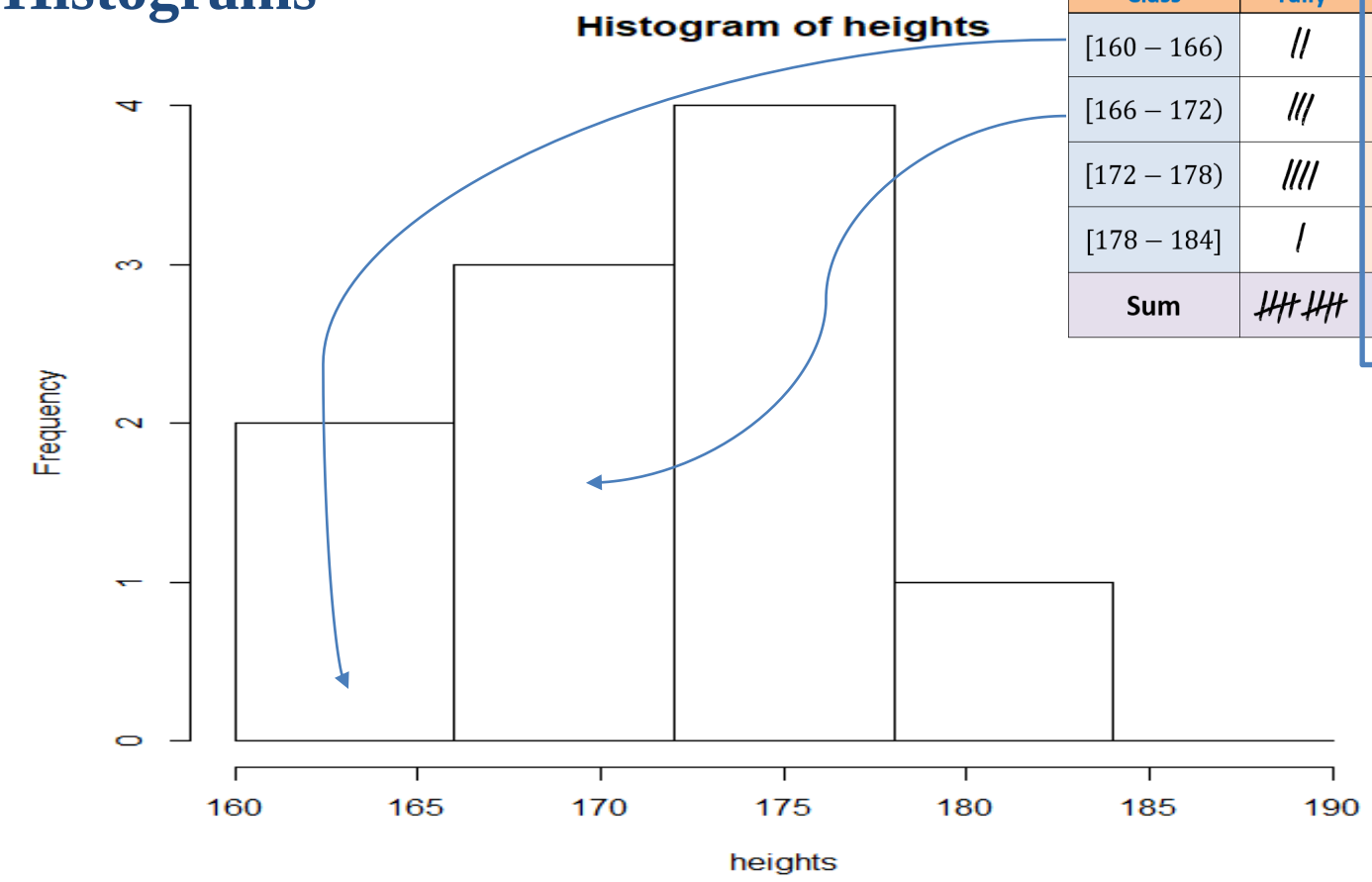
A histogram is a graphical tool consisting of bars placed side by side on a set of intervals (classes, bins, or cells) of equal width. The bars represent the frequency or relative frequency of classes. The height of each bar is proportional to the frequency or relative frequency of the corresponding class.

Representation of Data (6/8)

Example5: (7/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Histograms



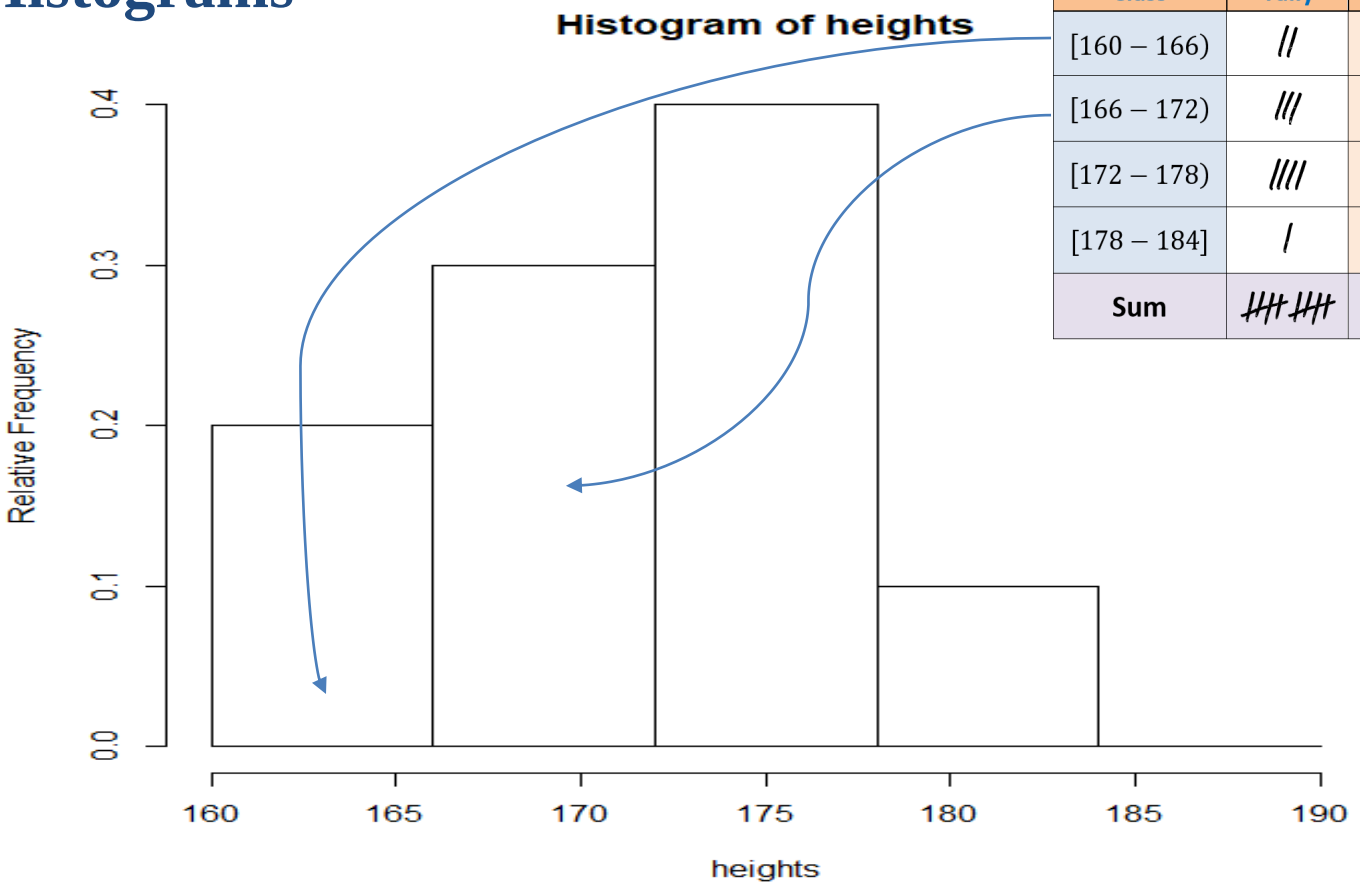
Class	Tally	Frequency (F)	Relative Frequency (R.F.)
[160 – 166)	//	2	2/10
[166 – 172)	///	3	3/10
[172 – 178)	////	4	4/10
[178 – 184]	/	1	1/10
Sum	~~//// ////~~	**10**	**1**

Representation of Data (6/8)

Example5: (8/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Histograms



Class	Tally	Frequency (F)	Relative Frequency (R.F.)
[160 – 166)	//	2	2/10
[166 – 172)	///	3	3/10
[172 – 178)	////	4	4/10
[178 – 184]	/	1	1/10
Sum	~~//// ////~~	**10**	**1**

Representation of Data (6/8)

Example5: (9/9)

175, 184, 160, 170, 173, 170, 175, 165, 171, 176

Example:

Using R

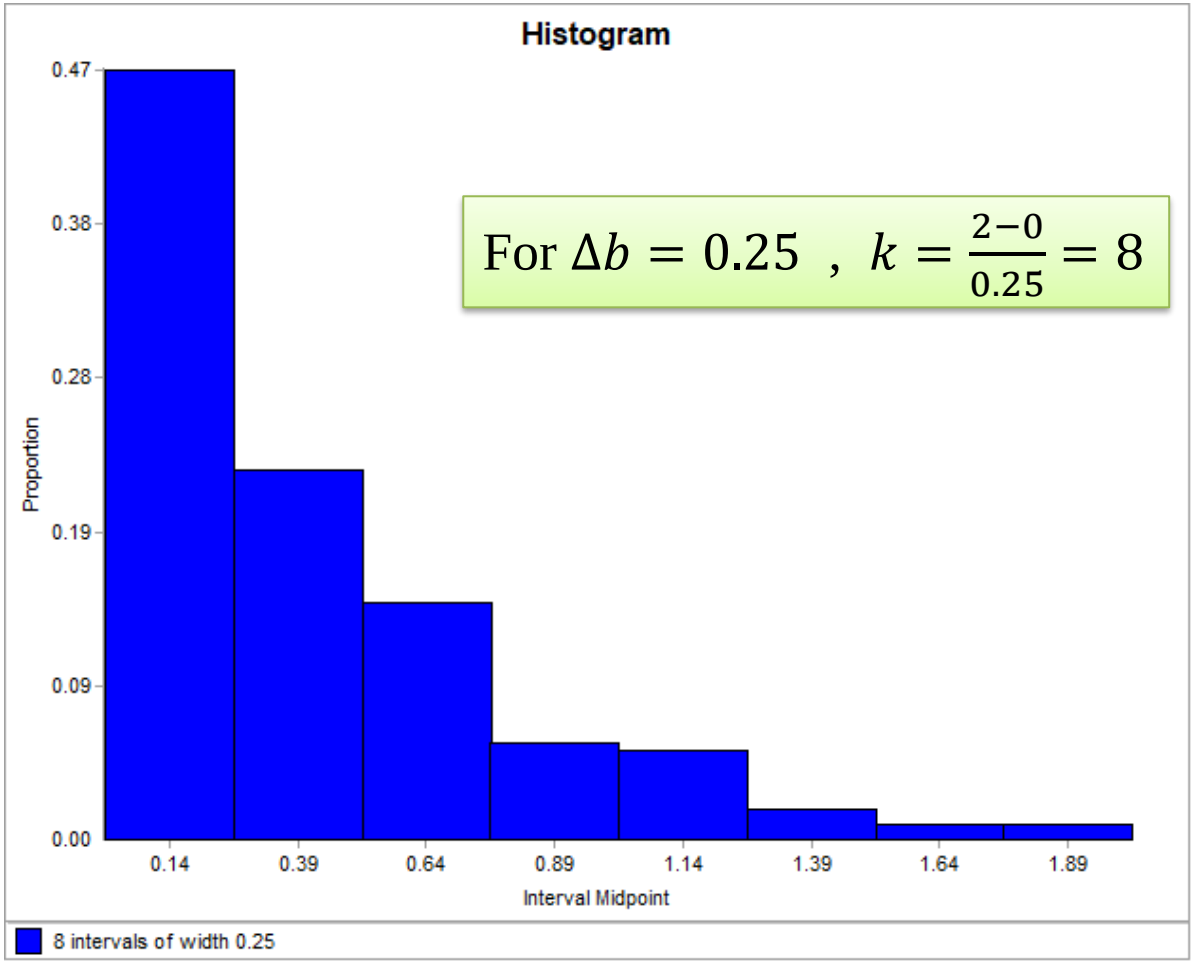
```
> heights <- c(175,184,160,170,173,170,165,175,171,176)

> hist(heights, breaks=seq(160, 190, by=6))

> install.packages("HistogramTools")
> library(HistogramTools)
> PlotRelativeFrequency(hist(heights, breaks=seq(160, 190, by=6)))
```

Representation of Data (7/8)

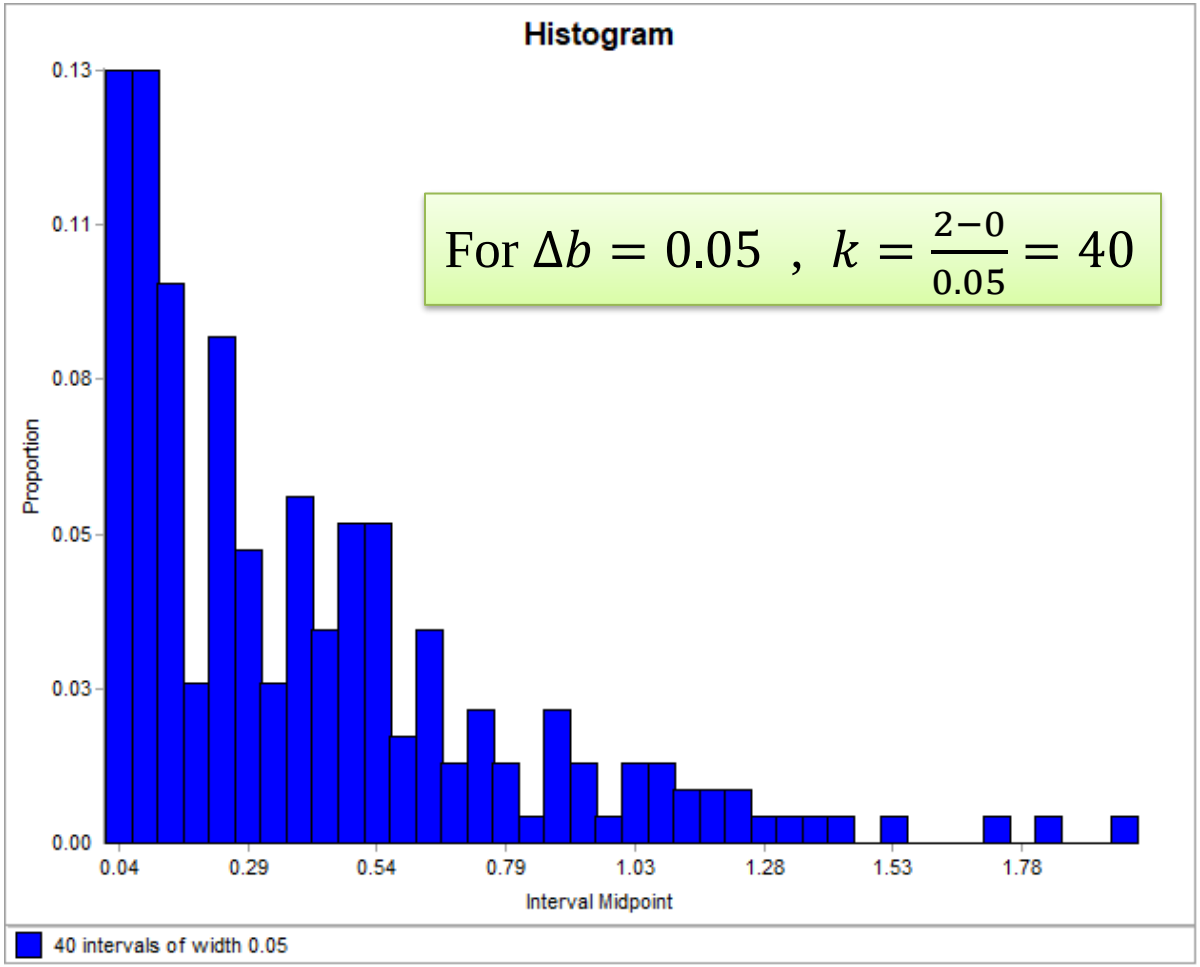
Histograms:



For $\Delta b = 0.25$, $k = \frac{2-0}{0.25} = 8$

Representation of Data (7/8)

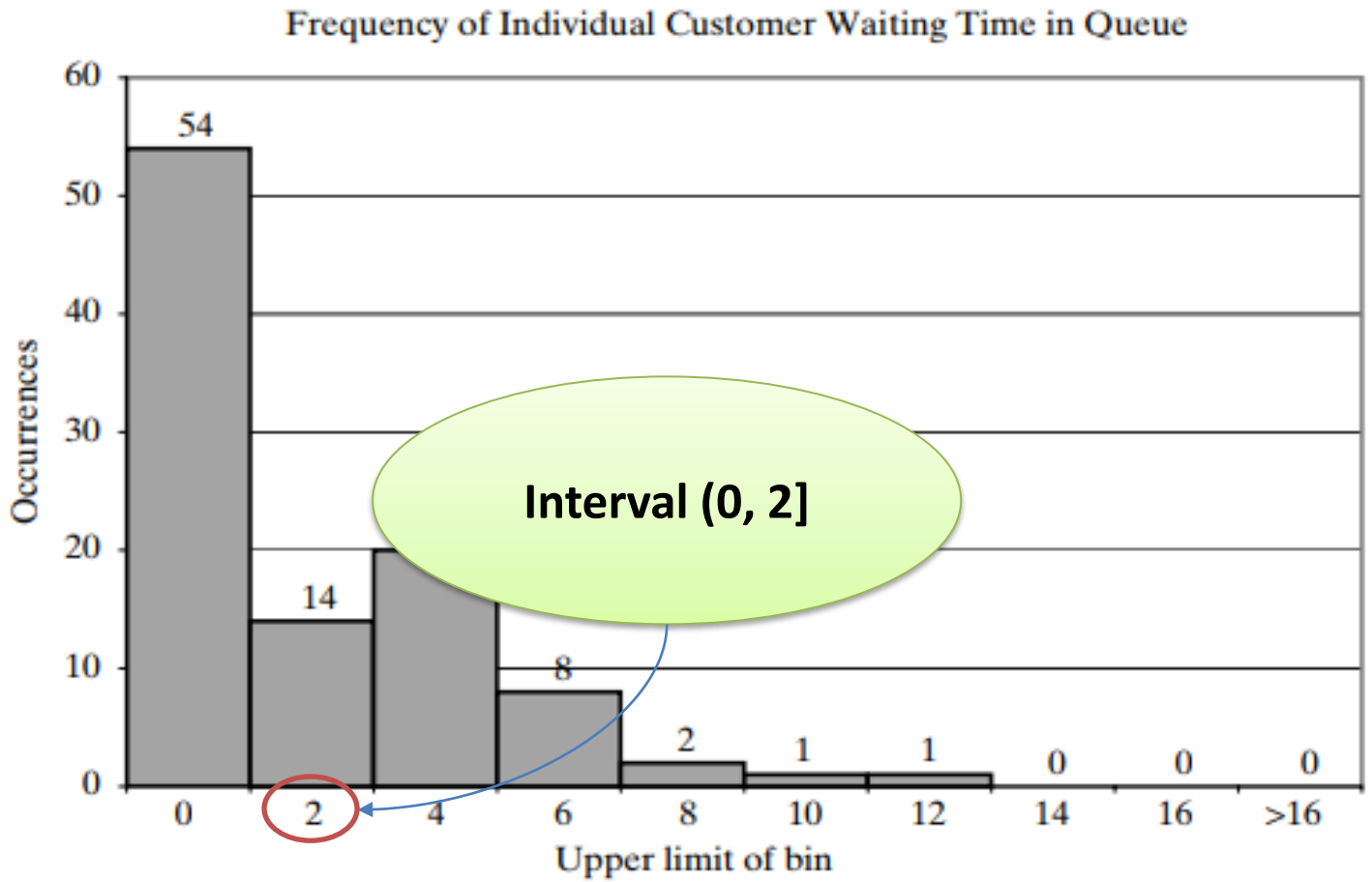
Histograms:



For $\Delta b = 0.05$, $k = \frac{2-0}{0.05} = 40$

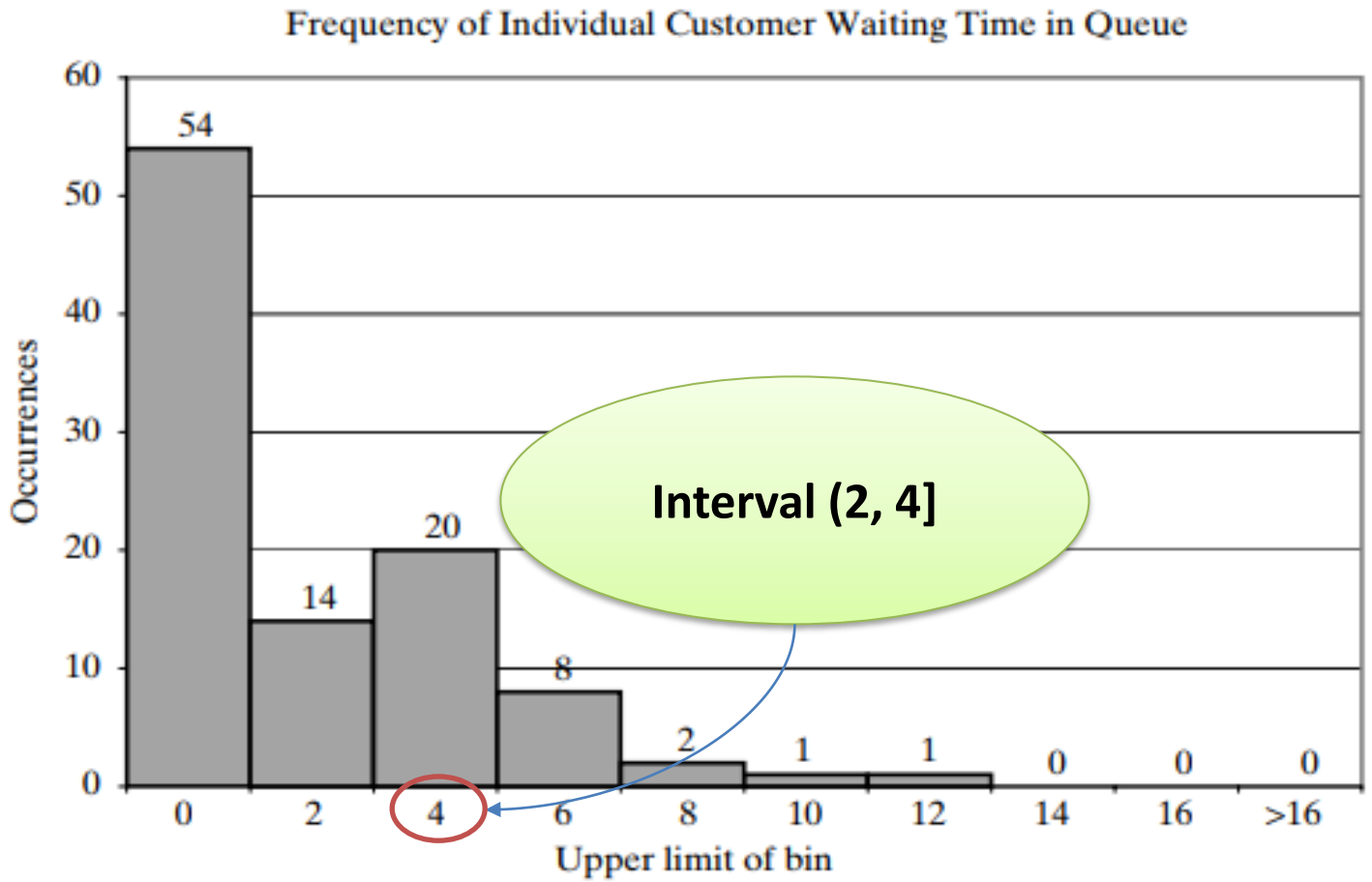
Representation of Data (8/8)

Histograms:



Representation of Data (8/8)

Histograms:



Revision Questions (1/2)

Given the following frequency table represent the colors represented in a sample:

Colors	Frequency
Red	5
Blue	8
Green	2

1. What is the type of data?
2. What is the sample size?
3. What graph can use to represent these data?
4. What is the most frequency colored?

Revision Questions (1/2)

Given the following frequency table represent the colors represented in a sample:

Colors	Frequency
Red	5
Blue	8
Green	2

1. What is the type of data? (Qualitative Data)
2. What is the sample size? = 15
3. What graph can use to represent these data? (Bar/Pie Charts)
4. What is the most frequency colored? “Blue”

Revision Questions (2/2)

Given the following table:

Classes	Frequency	Relative frequency
15 – 20	8	
20 – 25	10	
25 – 30	7	
30 – 35	5	

1. What is the type of data? ------
2. Complete the table
3. The sample size is ------
4. The table name is ------
5. The class interval = ------
6. The most frequency class = ------
7. The midpoint for the second class = ------
8. The name of graph represent this table is ------

Revision Questions (2/2)

Given the following table:

Classes	Frequency	Relative frequency
15 – 20	8	8/30
20 – 25	10	10/30
25 – 30	7	7/30
30 – 35	5	5/30

1. What is the type of data? (Continuous Quantitative Data)
2. Complete the table
3. The sample size = (30)
4. The table name is (Frequency Table)
5. The class interval = (5)
6. The most frequency class = [20 – 25)
7. The midpoint for the second class = (22.5)
8. The name of graph represent this table is (Histogram)

Video Lectures

All Lectures: https://www.youtube.com/playlist?list=PLxIvc-MGOs6gW9SgkmoxE5w9vQkID1_r-

Lecture #12: https://www.youtube.com/watch?v=9BP0kSrS0yc&list=PLxIvc-MGOs6gW9SgkmoxE5w9vQkID1_r-&index=14

Thank You

Dr. Ahmed Hagag
ahagag@fci.bu.edu.eg